Monika Kolvenbach/Wolfgang Teubert

DER MOLEX-GENERATOR: PROBLEME DER NOMINALFLEXION

1. Aufgaben und Funktionen des MOLEX-Generators

Wie in den Mitteilungen des IdS Nr. 7 S. 60 ff. beschrieben, wurde am IdS ein Programm entwickelt, der sog. MOLEX-Generator, das aus der Angabe eines Nomens, Verbs oder Adjektivs in der Normalform und der Angabe der zugehörigen Flexionsklasse sämtliche Vollformen (Formen mit Flexionsmorphemen) für dieses Wort zusammen mit den morphologischen Beschreibungen zu den generierten Vollformen erzeugt.

Dieser MOLEX-Generator, der Bestandteil der morphologischen Komponente des Systems PLIDIS der ehemaligen Abteilung LDV (vgl. dort) war, wurde von der Abteilung ZWD übernommen und wird in dieser Abteilung weiterhin eingesetzt. Haupteinsatzgebiet ist hier die Erstellung eines großen Vollformen-Wörterbuchs, das in Zukunft als Grundlage für die morpho-syntaktische Beschreibung von Texten dienen soll. Neben diesem Haupteinsatzgebiet wird der MOLEX-Generator dazu verwendet, für Belegaufsuchen in den Corpus-Texten die Vollformen zu generieren, um so gezielter suchen zu können und die Arbeit des Zusammenstellens und Aufschreibens der Vollformen zu ersetzen. Außerdem wird es - die angestrebte Vollständigkeit vorausgesetzt - über das große Vollformen-Lexikon möglich, jeder beliebigen Form innerhalb des Stichwortregisters die Grundform - bei Mehrdeutigkeiten: die möglichen Grundformen - zuzuordnen. Eine Rückführung auf eindeutige Grundformen wird erst nach Durchlaufen einer Syntaxanalyse möglich sein, da nur über die syntaktischen Zusammenhänge Mehrdeutigkeiten wie bei *süßen* (entweder Verb oder Adjektiv) vereindeutigt werden können. Erst mit der Syntaxanalyse kann die Fehlerquote entscheidend reduziert wer-

den.

Für die Verben, die Adjektive und die regelmäßigen Nomen arbeitet der MOLEX-Generator folgendermaßen: Aus der Normalform (bei Verben der Infinitiv, bei Adjektiven das unflektierte Adjektiv, bei Nomen der Nominativ Singular, soweit vorhanden, sonst der Nominativ Plural) wird die Matrix-Form bzw. werden die Matrix-Formen gebildet, an die die Flexive angehängt werden.

Zuerst noch einige Bemerkungen zum Begriff "Matrix-Form". Es wurde bewußt vermieden, hier einen Begriff wie 'Stammform' oder 'Grundform' zu verwenden, da diese Begriffe in der sprachwissenschaftlichen Terminologie mit bestimmten Inhalten versehen sind, die den hier angesprochenen Sachverhalt nur zum Teil treffen.

Matrix-Form ist im folgenden so definiert: "expandiert man ein Wort in seiner Normalform derart, daß man zu ihr alle Vollformen angibt,und reduziert man diese Vollformen um die Flexive, dann bildet der "Rest" die Matrix-Form".

Warum diese komplizierte Beschreibung notwendig wird, zeigen die folgenden Beispiele: Nimmt man z.B. das Verb *stellen*, dann erhält man nach Abstreichen der Flexionsendung "-en" für den Infinitiv die Matrix-Form *stell*, mit der sämtliche Vollformen gebildet werden können. Hier könnte man auch genauso gut sagen, man erhält den 'Stamm' *stell*. Nimmt man jedoch das Verb *herstellen* und streicht die Infinitiv-Endung *-en* ab, dann erhält man die Matrix-Form *her*stell*, mit der dann sämtliche Vollformen des Verbs *herstellen* gebildet werden können, sowohl in präfigierter als auch in unpräfigierter Form mit nachgestelltem Präfix. Der Stern '*' gewährleistet, daß in einem Fall wie *her*stellen* einmal alle Formen des Verbs in präfigierter Form für die Verwendung in Nebensätzen *(, weil die Fabrik Gummi herstellt)* und in unpräfigier-

ter Form für die Verwendung in Hauptsätzen (*Die Fabrik stellt Gummi her.*) generiert werden. Für die unpräfigierten Formen wird das abgetrennte Verbpräfix in der morphologischen Beschreibung unter VPR (Verbpräfix) angegeben, z.B. "(*stell* VERB ... VPR *her*))". Dem Sprachwissenschaftler würde sich der Filzschreiber sträuben, wenn er *her stell* als Stamm bezeichnet fände, wo doch die richtige Bezeichnung "Präfix + Stamm" wäre. Die Form *her stell* wäre jedoch unter diesem Aspekt immer noch als 'Grundform' akzeptabel, da sie aus einem Lexem in seiner Grundform nach Abstreichen der Infinitiv-Endung entstanden ist. Wenn man jedoch als Beispiel ein unregelmäßiges Verb, wie etwa *vertreiben* betrachtet, dann erhält man als Matrix-Form *vertreib* und *vertrieb*. Hier würde der Linguist nur *vertreib* als Grundform bezeichnen, nicht aber *vertrieb*. Man kann diese Formen jedoch wegen des Präfixes auch nicht als 'Stämme' bezeichnen.

Für den MOLEX-Generator kann man davon ausgehen, daß bei regelmäßigen Verben die Verben auf *-en* mit einer Matrix-Form auskommen, z.B.: *addier, arbeit*. Die regelmäßigen Verben auf *-n* benötigen jeweils zwei Matrix-Formen, nämlich *forder / fordr, lächel / lächl*. Wesentlich anders verhalten sich die unregelmäßigen Verben, die bis zu sechs Matrix-Formen benötigen, damit alle Vollformen generiert werden können, z.B.: *berst / barst / birst / borst / börst / bärst*. Diese Matrix-Formen können einserseits von Flexiven gefolgt werden, wie etwa in *borst-est*; Flexive können allerdings auch vorangestellt sein wie in *ge-borst-en*. Bei Matrix-Formen zu Verben mit abtrennbarem Präfix kann sogar innerhalb der Matrix-Form ein Flexiv eingefügt werden wie etwa bei *voran-treib*, aus der dann auch *voran-zu-treib-en* abzuleiten ist.

Nicht umlautende Adjektive kommen im Positiv, Komparativ und Superlativ mit einer Matrix-Form aus, wie etwa *schön, klein*. Zwei Matrix-Formen werden benötigt für die umlau-

tenden Adjektive, z.B. *groß / größ, dumm / dümm;* die unregelmäßigen Adjektive benötigen drei Matrix-Formen, z.B. *gut / bess / bes* bzw. vier *hoh / höh / höch* mit der Normalform *hoch*, die jedoch als ADJU (Adjektiv unflektiert) nicht flektieren kann.

Bei den Nomen kommen die regelmäßigen mit nur einer Matrix-Form aus, wenn sie nicht umlauten, z.B. *Kind, Tag.* Umlautende Nomen benötigen zwei Matrix-Formen, z.B. *Mann / Männ, Vogel / Vögel.*

2. Zur Generierung der Nomen

Unsere Flexionsklassenangabe bei den Nomen, die für die richtige Generierung unerläßlich ist, richtet sich für die regelmäßigen Nomen soweit wie möglich nach der Flexionsklasseneinteilung, die Gerhard Wahrig in seinem "Deutschen Wörterbuch" (Gütersloh 1968, S. 50-54) angegeben hat. Dies hat vor allem praktische Gründe: Da das MOLEX und der MOLEX-Generator nur als Hilfsmittel in einem größeren Projekt eingesetzt wurden, sollte der Aufwand möglichst gering gehalten werden. Deshalb würden die Nomen, Verben und Adjektive zur Erweiterung des morphologischen Lexikons entsprechend den Wahrig-Konventionen aufgenommen. Bei den Nomen wird (entsprechend dieser Konvention) nur das Simplex-Nomen mit einer Flexionsklassen-Angabe versehen, da in den meisten Fällen eine Zuordnung der Komposita zu den entsprechenden Flexionsklassen über die Genus-Angabe leicht durchführbar ist. Die Wahrig-Nomen-Flexions-Klassen enthalten darüberhinaus Angaben zum Umlaut für die Pluralformen, die deshalb keine weiteren Schwierigkeiten bilden. Einige Klassen, wie etwa die Klasse N11 müssen allerdings einer Sonderbehandlung unterzogen werden, da man nach der Vorgabe des Wahrig für *Erlebnis* die falsche Genitiv-Form *Erlebnises* und die falsche Pluralform *Erlebnise* erhalten würde. Grundsätzlich wird bei

allen Nomenklassen eine Abfrage eingebaut, ob die Matrix-Form auf *-s* auslautet. Z.B. werden über die Klasse M1 für *Tag* die unterschiedlichen Vollformen *Tag / Tages / Tags / Tage / Tagen* gebildet. Für das Lexem *Greis*, das ebenfalls der Klasse M1 angehört, dürfen jedoch nur die Vollformen *Greis / Greises / Greise / Greisen* gebildet werden, da die Vollform *Greis-s* bei Nomen dieser Klasse verboten ist.

Im Anhang finden sich einige Beispiele für die Nomen-Generierung, an denen auch die Beschreibungskategorien des MOLEX für Nomen-Einträge deutlich werden.

Dieses Verfahren wurde für regelmäßig flektierende Nomen (auch solche mit Umlaut) durchgehalten. Die unregelmäßigen Nomen sind im Wahrigschen Wörterbuch nicht in der Form klassifiziert, daß bei den Simplexformen (Nicht-Komposita-Formen) die Genus-Angabe samt der Flexionsklassen-Nummer steht, sondern bei ihnen wird zu der Simplex-Form die Endung für den Genitiv und die Nominativ-Plural-Form angegeben, z.B. *Jambus* (m.; Gen. -; Pl. *-ben*) (Sp. 1923). Teilweise sind auch mehrere Genitiv-Formen angegeben, z.B.: *Christus* (m.; unz.; Gen. *-i* od. -) (Sp. 813). Daß es sich hierbei nicht um "Fremdwörter" handelt, sondern diese Unregelmäßigkeit auch deutsche Wörter treffen kann, sieht man an dem Eintrag für *Bau* II (Sp. 576) "*Bau* (m.; Gen. *-(e)s*; Pl. *-ten*". Hierher gehören auch die Nomen, bei denen die Pluralformen von der Singularform nicht ableitbar sind, wie etwa *Edelmann - Edelleute, Grünland - Grünländereien.*

Für alle oben kurz skizzierten unregelmäßigen Fälle mußte nun ein Verfahren zur Generierung der Pluralformen gefunden werden, das zwei Bedingungen erfüllen sollte:

1. es sollte sich ohne Schwierigkeiten in das Verfahren für die regelmäßigen Nomen integrieren lassen und sollte mit diesem gemeinsam ablaufen, damit alle Nomen in

einem Lauf generiert werden konnten;

2. es sollte aus Speicherplatzgründen möglichst unaufwendig sein bezüglich der Angaben, die als Grundlage für die Generierung zur Verfügung gestellt werden mußten.

Man entschied sich deshalb für das im folgenden beschriebene Vorgehen:

Zusätzlich zu den im Wahrig vorhandenen Nomenklassen wurden für diese Fälle neue Klassen definiert, die folgende Eigenschaften haben: Über die Flexionsklassenangabe wird wie bei regelmäßigen Nomen die Generierung der Singular- und Pluralformen geleistet. Zusätzlich wird bei den unregelmäßigen Nomen dem Generator ein Hilfsmittel angeboten, mit dem er die zutreffenden Matrix-Formen leicht ermitteln kann. Dieses Hilfsmittel zur Ermittlung der Matrix-Formen gibt an, wie viele Buchstaben von rückwärts von der Normalform abgestrichen werden müssen, und durch welche Buchstaben diese ersetzt werden müssen, damit die richtige Plural-Matrix-Form entsteht. Als Beispiel sei *Basis* genannt. Hier bleibt die Form *Basis* für alle vier Singular-Kasus gleich, für den Plural muß *Basen* für alle vier Kasus erzeugt werden. Dies geschieht unter formalem Aspekt am besten dadurch, daß man von der Singularform die beiden letzten Buchstaben abstreicht und durch *-en* ersetzt. So wird dies auch notiert: (*Basis* (F 60 (-2 +*en*))). Die Klasse F 60 gewährleistet, daß die Singularform *Basis* für alle vier Kasus mit der entsprechenden Beschreibung generiert wird, ebenso wie die Form *Basen* als Pluralform, ebenfalls mit sämtlichen morphologischen Beschreibungen für die vier Kasus.

Als Sonderfälle sind die oben angegebenen Lexeme *Edelmann, Grünland, Bau* zu betrachten. Auch hier wurde nicht versucht, über Komposita-Regeln die Fälle herauszufinden, in denen z.B. der Komposita-Bestandteil *-mann* als *-leute*

im Plural auftritt bzw. wo das nicht der Fall ist. *Amtmann*, das den Plural ganz regelmäßig als *Amtmänner* bildet, wurde schließlich als "M 2U" gekennzeichnet, bei *Edelmann* wird die Matrix-Form zur Pluralbildung durch die Angabe (-4 *+leute*) erzeugt, die Flexionsklasse (M 51) sorgt dafür, daß sowohl sämtliche Singularformen als auch die beiden Pluralformen gebildet werden. Das gleiche gilt für *Grünland*, bei dem die Pluralform ebenfalls nach obigem Schema durch (-4 *+ländereien*) erzeugt wird. Ebenso wurde für *Bau*, obwohl es in Zusammensetzungen häufig auftritt, keine eigene "regelmäßige" Klasse gebildet, sondern es wird ebenso wie die anderen Ausnahmen behandelt und seine zweite Matrix-Form wird über (-O *+ten*) erzeugt.

Hierbei stand einzig und allein die möglichst große Effektivität im Vordergrund sowie die einfache Behandlung innerhalb eines durch die regelmäßigen Nomen vorgegebenen Verfahrens.

Nehmen wir nun als Beispiel das Nomen *Klassizismus*, das ebenfalls im Wahrig nicht klassifiziert ist, und stellen hieran einige Betrachtungen darüber an, inwieweit dieses Verfahren zur Gewinnung der Matrix-Form gerechtfertigt ist. Der Begriff Matrix-Form wird sehr praxisorientiert verwendet. Im obigen Beispiel dürfte eigentlich *Klassizismus* nicht als Matrix-Form verwendet werden, weil die eigentliche Matrix-Form aus diachronischer Sicht *Klassizism-* wäre, an die im Singular das Singular-Flexiv *-us* angehängt wird, das im Plural durch das Flexiv *-en* ersetzt wird. Da jedoch entsprechend unseren Vorgaben nur ein Verfahren mit möglichst gleichförmigem Aufbau angestrebt wurde, mußten die linguistischen Unzulänglichkeiten in Kauf genommen werden, um wenigstens formal "sauber" arbeiten zu können, zumal es nur darauf ankam, alle Vollformen mit ihren Beschreibungen zur Verfügung zu stellen. Somit hat das eigentliche Generierungsverfahren also nur Hilfsmittel-Funktion. Linguistische Theorien waren nur

insofern einzubeziehen, als sie einen Beitrag zu einem vollständigen, zufriedenstellenden Ergebnis lieferten. Daher notieren wir bei *Klassizismus (Klassizismus* (M 50 (-2 *+en*))).

An diesem Beispiel ist jedoch eine andere Defizienz des Morphologischen Lexikons insgesamt aufzuzeigen: In dem über die Genus- und Flexionsklassenangabe zu den einzelnen Wörtern erzeugten Vollformenlexikon ist es erforderlich, die Vollformen in der Gesamtheit ihres möglichen Vorkommens aufzunehmen. Das bedeutet, daß, wenn Singular und Pluralformen zu einem Lexem vorhanden sind, diese auch erzeugt werden und an die Normalform angebunden werden. Für das obige Beispiel bedeutet das: es wird kein Unterschied zwischen 'Klassizismus' als Epoche und 'Klassizismus' (als Stilmittel) gemacht. In der ersten Bedeutung wäre nach deutschem Sprachgebrauch nur der Singular anzusetzen, in der zweiten Bedeutung sowohl Singular als auch Plural.

Um in dem Vollformenlexikon jedoch alle Vollformen, die von einer Normalform abgeleitet werden können, zu erfassen, ist, da man in diesem speziellen Lexikon keine semantischen Informationen unterbringen kann und dies auch nicht Aufgabe des Lexikons ist, die semantische Unterscheidung irrelevant. Entscheidend sind also nur die formalflexionsmäßigen Möglichkeiten in möglichst umfassender Form. Alle anderen Unterscheidungen haben in diesem Lexikon keinen Platz.

Die semantische Disambiguierung setzt in einer späteren Phase ein. Erst wenn aufbauend auf der vollständigen morphologischen Beschreibung eines jeden Einzelwortes im Satz die syntaktischen Zusammenhänge innerhalb des Satzes mittels einer Syntaxanalyse analysiert sind, kann die Übersetzung in eine Semantiksprache erfolgen, die dann eben über die semantiksprachlichen Regeln diese Unterschiede aufdeckt und verarbeitet. Bei dieser Übersetzung wird dann über syntax-sensitive Regeln und Umgebungsanalysen eine

adäquate semantische Beschreibung des Satzes geliefert, in der Unterscheidungen bezüglich der Verwendungsweise von Homographen, wie etwa *Klassizismus* als 'Epoche' und als 'Stilmittel' bei genügend verfeinertem Beschreibungsapparat und entsprechender Ausrichtung der Zielsetzung erkannt und sichtbar gemacht werden können. Ein derart verfeinerter Beschreibungsapparat sollte nicht nur in der Lage sein, sich bei Vorliegen der Form *Klassizismen* eindeutig für die Lesart 'Stilmittel' zu entscheiden, er sollte so konzipiert sein, daß er die beiden Verwendungsformen aufgrund der Umgebungsanalyse auch im Singular zu unterscheiden in der Lage ist.

Hierzu muß neben einem Übersetzungsalgorithmus in eine geeignete Semantiksprache ein Lexikon vorhanden sein, das die Übersetzungsregeln für die verschiedenen Bedeutungen eines Lexems in dem gewählten Anwendungsgebiet in ähnlicher Vollständigkeit enthält, wie sie für das MOLEX angestrebt wurde. Da das MOLEX jedoch nur für die Morphologie konzipiert wurde, muß die "Vollständigkeit" dieses semantiksprachlichen Lexikons anders definiert werden, nämlich als "Vollständigkeit in einem bestimmten Anwendungsausschnitt der Sprache". Nur so kann man zu einer sinnvollen Verwendung der Semantiksprache gelangen, da eine vollständige semantiksprachliche Beschreibung von Sprache ohne Bezug auf ein eingegrenztes Anwendungsgebiet nicht leistbar ist.

Wer sich für die einzelnen Schritte von einem natürlichsprachlichen Eingabesatz über die morphologische Beschreibung der Einzelwörter in diesem Satz, daran anschließend das Zusammenfügen der Vielzahl der morphologischen Beschreibungen zu einer syntaktischen Beschreibung der Satz-Entitäten und schließlich zu einer streng sachgebietsorientierten semantiksprachlichen Darstellung der Satzinhalte interessiert, sei auf die PLIDIS-Dokumentation verwiesen, in der auf der Grundlage eines ablauffähigen Informationssystems (zur Abwasserüberwachung) das Ineinandergreifen der einzelnen oben

skizzierten Schritte dargestellt ist.

Dokumentation zu PLIDIS als Testversion:

Lutz, Hans Dieter
Kurzdokumentation über das System PLIDIS, Version 2.0, Oktober 1977.
Mannheim, Institut für deutsche Sprache. April 1980.
(56 Seiten DIN A4, DM 10,--)

Dokumentation zu PLIDIS als experimentelles Anwendungssystem:

Lutz, Hans Dieter; Kolvenbach, Monika; Zifonun, Gisela; et al.
PLIDIS Dokumentation.
Mannheim: Institut für deutsche Sprache. 1980.
(ca. 500 Seiten DIN A4, DM 68,--)

Beide Dokumentationen sind zu beziehen über:

Institut für deutsche Sprache
Abteilung Zentrale Wissenschaftliche Dienste
Postfach 5409
D-6800 Mannheim 1

Anhang

Beispiele für die Nomengenerierung

a) Eingabe-Datei (Nomen in der Normalform + Klasse)

```
(BASIS (F 51 (-2 +EN)))
(BASE (F 19))
(SANDFANG (M 1U))
(CYANID (N 11))
```

b) Ausgabedatei (generierte Vollformen + morphologische Beschreibung)

```
(BASE (N NF BASE KNG 7746 K (NOM GEN DAT AKK) PN 3 G F))
(BASEN 2 (N NF BASIS KNG 7690 K (NOM GEN DAT AKK) PN 6 G F))
       1 (N NF BASE KNG 7690 K (NOM GEN DAT AKK) PN 6 G F))
(BASIS (N NF BASIS KNG 7746 K (NOM GEN DAT AKK) PN 3 G F))
```

```
(CYANID (N NF CYANID KNG 5697 K (NOM DAT AKK) PN 3 G N))
(CYANIDE (N NF CYANID KNG 6665 K (NOM GEN AKK) PN 6 G N))
(CYANIDEN (N NF CYANID KNG 1033 K DAT PN 6 G N))
(CYANIDES (N NF CYANID KNG 2113 K GEN PN 3 G N))
(CYANIDS (N NF CYANID KNG 2113 K GEN PN 3 G N))

(SANDFAENGE
        (N NF SANDFANG KNG 6668 K (NOM GEN AKK) PN 6 G M))
(SANDFAENGEN
        (N NF SANDFANG KNG 1036 K DAT PN 6 G M))
(SANDFANG
        (N NF SANDFANG KNG 5700 K (NOM DAT AKK) PN 3 G M))
(SANDFANGE
        (N NF SANDFANG KNG 1092 K DAT PN 3 G M))
(SANDFANGES
        (N NF SANDFANG KNG 2116 K GEN PN 3 G M))
(SANDFANGS
        (N NF SANDFANG KNG 2116 K GEN PN 3 G M))
```

c) Verwendete Kategorien und Abkürzungen

N = Nomen (vor NF)
NF = Normalform
KNG = Kasus, Numerus, Genus in verschlüsselter Form (vgl. hierzu PLIDIS-Dokumentation)
K = Kasus
NOM GEN DAT AKK = Nominativ, Genitiv, Dativ, Akkusativ
PN = Person/Numerus
PN 3 = 3. Person Singular
PN 6 = 3. Person Plural
G = Genus
M F N = Masculin, Feminin, Neutrum in Verbindung mit G

Die Unterscheidung in Groß- und Kleinbuchstaben entfällt, da die Großschreibung bei der Syntax-Analyse nicht als Kriterium verwendet wird. Umlaute werden durch nachgestelltes "E" gekennzeichnet, ß wird als "SS" dargestellt.

d) Bemerkungen

Ordnungskriterien für die Einträge ist die Vollform. Identische Vollformen, die von unterschiedlichen Normalformen abgeleitet werden, werden in einen Eintrag zusammengefaßt (vgl. BASEN). Das gleiche gilt für Vollformen, die auf unterschiedliche Wortarten zurückgehen, z.B. auf Verben, Nomen und Adjektive.

Die Angabe zu den Kategorien Kasus, Numerus und Genus ist bei Nomen obligatorisch. Fakultative Kategorien-Angaben gibt es nicht. (Zu den Kategorien der anderen Wortarten vgl. die PLIDIS-Dokumentation).